Mean Normalized Scores by Model and Category

| Model | Administration and Workflow | Clinical Decision Support | Clinical Note Generation | Medical Research Assistance | Patient Communication and Education |
|---|---|---|---|---|---|
| DeepSeek R1 | 0.61 | 0.70 | 0.85 | 0.70 | 0.89 |
| o3-mini (2025-01-31) | 0.61 | 0.77 | 0.85 | 0.75 | 0.84 |
| Claude 3.5 Sonnet (20241022) | 0.63 | 0.69 | 0.82 | 0.71 | 0.83 |
| Claude 3.7 Sonnet (20250219) | 0.63 | 0.70 | 0.83 | 0.66 | 0.84 |
| GPT-4o (2024-05-13) | 0.62 | 0.71 | 0.80 | 0.72 | 0.82 |
| Gemini 2.0 Flash | 0.56 | 0.68 | 0.78 | 0.68 | 0.81 |
| GPT-4o mini (2024-07-18) | 0.61 | 0.68 | 0.79 | 0.66 | 0.81 |
| Llama 3.3 Instruct (70B) | 0.53 | 0.68 | 0.76 | 0.65 | 0.81 |
| Gemini 1.5 Pro (001) | 0.57 | 0.63 | 0.74 | 0.68 | 0.76 |